

REMARKS

The Final Office Action included new statements that addressed arguments made in Applicants' previous response. Applicants will address these statements first.

In the Final Office Action, the Examiner stated that the arguments with regards to the fact that Smith does not make a selection between an orthographically derived acoustic description and a speech-based acoustic description based on the scores for the acoustic descriptions was not true as evidenced by column 12, lines 30-37 of Smith. However, the cited section does not state that a selection is being made between a speech-based acoustic description and an orthographically derived acoustic description. Instead, it only describes text-based acoustic descriptions. As such, Applicants maintain the position that Smith does not show the selection between a speech-based acoustic description and an orthographically derived acoustic description.

The Final Office Action also stated that Applicants' argument that Bahl '921 does not generate scores for syllable-like units is contradicted with the establishing of syllables as a basic unit to be considered with the referenced phones. In the previous response, Applicants made two arguments with regards to Bahl '921. First, Applicants argued that Bahl '921 does not show or suggest that its language model uses syllable-like units. Even if a syllable is used as the basic acoustic unit for the acoustic models, there is no suggestion in Bahl '921 that the syllable should be used in the language model. Second, Applicants have argued that Bahl '921 does not show or suggest generating an acoustic model score for a sequence of syllable-like units by generating acoustic model scores for each of the sequence of phonemes that form the sequence of syllable-like units. Under Bahl '921, if an acoustic model score is determined for a syllable, it is determined for the entire syllable in one step. It is not determined by determining the acoustic model

scores for phonemes that form the syllable-like unit. This can be seen in column 2, lines 31-35 where Bahl '921 states that "The words in the transcription are assumed to be expressed in terms of certain basic units or classes such as phonemes, syllables, words or phrases and the acoustic model is essentially composed of models for each of these different units." Thus, if Bahl '921 uses syllables as the basic units, then the acoustic model is syllable-based, and not phoneme-based. As such, the acoustic model score for a syllable in Bahl '921 would not be formed by generating an acoustic model score for a sequence of phonemes that form the syllable-like unit but instead would be formed directly from a syllable model.

In response to Applicants' argument that Cantolini does not show generating an acoustic model score for a sequence of syllable-like units by generating acoustic model scores for each of a sequence of phonemes that form the sequence of syllable-like units, the Examiner cited column 6, lines 56-65 and stated that the syllable-like characteristic comprises the leaf nodes of the tree and that the score estimator rescores each.

The leaf nodes referred to in Cantolini each represent a different phoneme. However, Cantolini does not show or suggest that a sequence of scores of leaf nodes are used to form an acoustic model score for a syllable-like unit. Further, the cited section does not involve acoustic model scores, but instead discusses rule-based scoring that generates a probability of a phoneme being produced from a letter. Thus, it is an orthographically derived probability not an acoustic model score.

With regard to Applicants' argument that Gupta does not produce a speech-based phonetic description, the Final Office Action stated that the cited reference does not support this argument. In particular, column 5, lines 24-39 of Gupta were said to counter Applicants' argument that if all of the text-based phonetic descriptions were removed from Gupta, generator

400 would produce an empty graph. In making this statement, the Examiner asserted that the phoneme rules of Gupta generate possible phonetic transcriptions. Applicants note that the cited phoneme rules in Gupta are letter-to-phoneme rules. Thus, these rules require letters and therefore are text-based and not speech-based phonetic descriptions. As such, Applicants maintain the position that Gupta does not show or suggest producing speech-based phonetic descriptions. Instead, Gupta only produces text-based phonetic descriptions.

The Final Office Action also stated that "With regard to claims 19-21, the citation of Schulze is found to apply only to an exception to the invention, those languages that cannot be divided into words or syllable-like units and so the reference is seen to be out of context to the argument." Regretfully, Applicants do not understand this statement. As such, they are unable to address it in a meaningful way.

Lastly, the Final Office Action asserted that the disclaimer at col. 16 lines 37-39 of Schulze showed that there was latitude in how trigrams are identified in a word. Applicants apologize for missing this disclaimer and agree with the Examiner that there is some latitude in how trigrams are identified in Schulze. However, Applicants still assert that Schulze does not show breaking a word into speech units by preferring some speech units over others. In particular, Schulze does not provide any teaching for giving a preference to a particular trigram while breaking a word into trigrams. It is only after Schulze has broken a word into trigrams that Schulze determines whether to maintain the trigrams in a trigram array. Thus, Schulze breaks the word into trigrams before any preference is given to a particular trigram.

The final Office Action also repeated the arguments made in the Office Action of December 9, 2003. Applicants' responses to those arguments remain the same with the exception

of the arguments for claims 19-21. The arguments concerning claims 19-21 have been changed in light of the fact that Schulze does provide some latitude in identifying trigrams and the fact that the Examiner did not maintain some of the reasons for rejecting claims 19-21 in the Final Office Action. Applicants' responses to those arguments are repeated below with appropriate changes to the arguments concerning claims 19-21.

Claims 1-4

Claims 1-4 were rejected under 35 U.S.C. § 103(a) as being unpatentable over Smith et al. (U.S. Patent Number 6,408,271 B, hereinafter Smith) in view of Häb-Umbach et al. (U.S. Patent Number 5,873,061, hereinafter Häb-Umbach).

Smith discloses a system for generating possible pronunciations of a sequence of words. Under Smith, each word has many possible pronunciations. As a result, for a sequence of words, there are multiple possible combinations of these pronunciations. Smith selects the top N pronunciations for the sequence of words to store in a dictionary and to use during speech recognition. During speech recognition, a decoder compares input feature vectors to pronunciations in the dictionary to determine if any of the pronunciations match the user's speech.

Häb-Umbach identifies sub-word units for a new word by averaging a plurality of utterances of the new word into a reference template. The reference template is then compared against stored phonetic models to select the sub-word units that most likely produced the reference template. Häb-Umbach also includes a grapheme-to-phoneme conversion that converts text into sub-word units. Häb-Umbach, however, does not score the sub-word units produced by the grapheme-to-phoneme conversion and does not select between the phoneme sequence produced by the grapheme-to-phoneme conversion and the phoneme sequence formed from the

reference template based on a score for the grapheme-to-phoneme sequence.

Independent claim 1 provides a method of adding an acoustic description of a word to a speech recognition lexicon. Initially, the text of the word is converted into an orthographically derived acoustic description of the word. The orthographically derived acoustic description is then scored based in part on a comparison between the orthographically derived acoustic description and a speech signal representing a user's pronunciation of the word. The speech signal is also used to identify a speech-based acoustic description of the word and a score for the speech-based acoustic description wherein the speech-based acoustic description is not associated with the text of the word. One of the orthographically derived acoustic description and the speech-based acoustic description is then selected as the acoustic description of the word based on the scores for the two acoustic descriptions.

The combination of Smith and Häb-Umbach does not show or suggest the invention of claim 1 because neither reference shows or suggests selecting one of an orthographically derived acoustic description and a speech-based description based on the scores for the two acoustic descriptions.

In the Office Action, it is asserted that column 12, lines 26-37 of Smith showed a step of selecting one of an orthographically derived acoustic description and a speech-based acoustic description based on the scores for these acoustic descriptions. Applicants respectfully dispute this assertion.

The cited section discusses generating possible pronunciations for a sequence of words based on possible pronunciations for individual words in the sequence. All of the pronunciations are generated from text. As such, Smith can not take into consideration a score for a speech-based acoustic description that is not associated with the text of the word, but

instead can only score acoustic descriptions that are based on text.

Similarly, Häb-Umbach does not show or suggest selecting one of an orthographically derived acoustic description and a speech-based acoustic description based on a score for the orthographically derived acoustic description and a score for the speech-based acoustic description.

Since neither Smith nor Häb-Umbach show a step of making a selection based on the score of an orthographically derived acoustic description and the score of a speech-based description, their combination does not form the invention of claim 1. As such, claim 1 is not obvious from the combination of Smith and Häb-Umbach.

Claim 5

Claim 5 was rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Häb-Umbach and further in view of Bahl et al. (U.S. Patent Number 5,875,426 hereinafter Bahl '426). Bahl '426 provides a speech recognition system that is able to handle word pronunciations that are context dependent. During recognition, Bahl '426 first considers all possible stored pronunciations for all words in a vocabulary. The speech signal is applied to these pronunciations to identify a set of candidate words. All of these pronunciations are associated with the text of the words. These candidate words are applied to a language model that generates a score for each current candidate word based on a previously identified word. This results in a ranked list of candidate current words and the dictionary-based pronunciations of those words.

Bahl '426 then examines a field in each current word's dictionary entry and a field in the preceding word's dictionary entry to determine if an additional pronunciation of the word should be added as a candidate. Note that this additional pronunciation candidate is a rule-based candidate associated with

the text of the word and is not dependent on how the speaker pronounced the word. The speech signal is then applied to these candidate words and pronunciations in order to select a most likely word.

Dependent claim 5 depends from claim 1 and includes a further limitation wherein identifying a score for a speech-based acoustic description further comprises using a language model. Because claim 5 depends from claim 1, it includes the limitation to selecting one of the orthographically derived acoustic description and the speech-based acoustic description based on the score for the orthographically derived acoustic description and the score for the speech-based acoustic description. None of Smith, Häb-Umbach, or Bahl '426 show such a limitation.

In particular, Bahl '426 does not show such limitation, since it does not generate a speech-based acoustic description that is not associated with the text of the word. Since it does not generate such a speech-based acoustic description, it cannot score a speech-based acoustic description and as such cannot select one of an orthographically derived acoustic description and a speech-based acoustic description based on a score for a speech-based acoustic description.

Since none of Smith, Häb-Umbach, and Bahl '426 show selecting one of a speech based acoustic description and an orthographically derived acoustic description based on scores for the acoustic descriptions, claim 5 is patentable over the combination of Smith, Häb-Umbach, and Bahl '426.

Claim 6-8

Claims 6-8 were rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Häb-Umbach and Bahl '426 and further in view of Bahl et al. (U.S. Patent Number 6,377,921 hereinafter Bahl '921).

Bahl '921 provides a system for identifying transcription errors in text used for training a speech

recognition system. Bahl '921 trains a set of acoustic models for acoustic units such as words, syllables, and phones. After the training is complete, a speech signal is aligned with its corresponding transcript using the trained models and a score is determined for each acoustic unit in the transcript. Instances of acoustic units that receive a low score from these models are then flagged and examined by a human operator to determine if the transcription is in error.

Claims 6-8 depend indirectly from claim 1. As a result, they include the limitation to selecting one of an orthographically derived acoustic description and a speech-based acoustic description based on the score for the orthographically derived acoustic description and the score for the speech-based acoustic description. The combination of Smith, Häb-Umbach, Bahl '426, and Bahl '921 does not show or suggest this limitation.

As discussed above, Smith, Häb-Umbach, and Bahl '426 fail to show a selection based on both a score for an orthographically derived acoustic description and a score for a speech-based acoustic description. Similarly, Bahl '921 fails to show this limitation. Under Bahl '921, speech signals apply to a known transcription of the speech that is associated with the text of the words. As such, Bahl '921 does not score a speech-based acoustic description as found in claim 1. Therefore, it can not select an acoustic description based on a score for a speech-based acoustic description. Thus, none of the cited references show a selection of an acoustic description based on both a score for an orthographically derived acoustic description and a score for a speech-based acoustic description.

In addition, in claim 6, generating a score for a speech-based acoustic description includes generating a language model score for a sequence of syllable-like units. None of Smith, Häb-Umbach, Bahl '426, or Bahl '921 show or suggest

generating a language model score for a sequence of syllable-like units.

In the Office Action, language model 18B of Bahl '921 was cited as providing a language model score for syllable-like units. However, Bahl '921 never states that the language model uses syllable-like units. As such, it does not show or suggest generating a language model score for a sequence of syllable-like units.

Since none of the cited references show or suggest generating a language model score for a sequence of syllable-like units and since none of the cited references select an acoustic description based on a score for an orthographically derived acoustic description and a score for a speech-based acoustic description, claim 6 and claims 7 and 8, which depend therefrom, are patentable over Smith, Häb-Umbach, Bahl '426, and Bahl '921.

Claims 9-11

Claims 9-11 are rejected under 35 U.S.C. § 103(a) as being unpatentable over Smith in view of Häb-Umbach, and Bahl '426, and further in view of Contolini et al. (U.S. Patent Number 6,233,553, hereinafter Contolini).

Contolini provides a method of selecting one pronunciation from a set of text-based pronunciations. Under Contolini, a plurality of text-based pronunciations are formed from the spelling of a word using a transcription generator. The top N pronunciations are provided to a speech recognition system, which applies a speech signal to the transcriptions representing each pronunciation. The transcription that scores highest is selected for storage. Contolini does not identify a speech-based acoustic description from a speech signal where the speech-based acoustic description is not associated with the text of a word, nor does it show the production of an acoustic model score for a syllable-like unit by generating acoustic model scores for each of a sequence of phonemes that form the syllable-like unit.

Claims 9-11 depend from claim 1 and as such include the limitation to selecting an acoustic description based on both a score for an orthographically derived acoustic description and a speech-based acoustic description. None of the cited references show or suggest this limitation. In Contolini, a speech signal is applied against previously identified transcriptions to identify a score for each transcription. Since each of these transcriptions is associated with the text of the word, Contolini does not identify a speech-based acoustic description that is not associated with the text of a word. As such, Contolini can not select an acoustic description based on a score for a speech-based acoustic description.

Since none of the cited references show a step of selecting an acoustic description based on a score for both an orthographically derived acoustic description and a speech-based description, claims 9-11 are patentable over the cited art.

In addition, none of the cited references show or suggest generating an acoustic model score for a sequence of syllable-like units by generating acoustic model scores for each of a sequence of phonemes that form the sequence of syllable-like units as found in claim 9.

In the Office Action, it was asserted that claim 4, column 7, line 6 and column 6, line 56 of Contolini show this limitation. Applicants respectfully dispute this assertion.

Claim 4 simply states that the sound units of claim 1 are acoustic units. Neither claim 1 nor claim 4 make any mention of syllable-like units or of determining an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes that form the syllable-like units. Column 6, line 56 describes classes of phonemes including consonant and syllabic. This section does not suggest generating an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes. Instead, it simply shows that a single phoneme may

act as a syllable at times. When this occurs, forming an acoustic score for the syllable does not require determining the acoustic score for a sequence of phonemes. Instead, the acoustic score for a single phoneme is determined.

Column 7, line 6 discusses filtering unlikely sequences of phonemes. It does not show or suggest determining an acoustic score for a syllable-like unit by generating acoustic scores for each of a sequence of phonemes that form the syllable-like unit.

Since none of the cited references show or suggest determining an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes that form the syllable-like unit, the combination of these references does not show or suggest claim 9.

Claims 12-17

Claims 12-17 were rejected under 35 U.S.C. § 103(a) as being unpatentable over Gupta et al. (U.S. Patent Number 6,243,680, hereinafter Gupta) in view of Häb-Umbach.

Gupta provides a system for selecting a pronunciation of a word for entry into a dictionary. Under Gupta, the text of a new word is first converted into a string of phonemes using a set of text-to-phoneme rules 412. These phonemes are placed in a graph structure with each branch in the structure being represented by a different phoneme. For each phoneme branch, a set of parallel branches are constructed, one for each phoneme that is similar to the initial phoneme in the graph. Additional parallel branches are then added for each allophone of each phoneme in the graph where an allophone is a particular pronunciation of a phoneme. Gupta then applies a set of speech utterances to the graph to score each path through the graph. The path with the highest score is selected as the pronunciation of the word.

Independent claim 12 provides a computer-readable medium having instructions for selecting a phonetic description

of a word to add to a lexicon. These steps include receiving the text of the word and a speech signal representing a person's pronunciation of the word. The text of the word is converted into a text-based phonetic description while the speech signal is used to generate a speech-based phonetic description of the word without using the text of the word. Either the text-based phonetic description or the speech-based phonetic description is then selected for entry in the lexicon based on the correspondence between each phonetic description and the speech signal.

The combination of Gupta and Häb-Umbach does not show or suggest the invention of claim 12 because the combination does not include a step of selecting between a text-based phonetic description and a speech-based phonetic description.

In the background section of Gupta, different types of systems for identifying pronunciations of new words are described. In one system, an expert listens to the word and identifies the acoustic description. In a separate system, a continuous allophone recognizer is used that decodes speech utterances to identify an acoustic description that is not associated with a word. In another system, a set of text-based rules are used to form an acoustic description.

However, Gupta does not show or suggest determining an acoustic description from the text and an acoustic description from the speech signal and then selecting between the two acoustic descriptions. Instead, text-based acoustic descriptions are used in separate systems from speech-based acoustic descriptions.

Note that in the Gupta system itself, only text-based acoustic descriptions are used. Specifically, "[t]he feature vectors for each utterance are used to score the allophonic graph generated on the basis of the orthographic representation of the new word." (Gupta, col. 13, lines 61-63). Thus, graph scoring

unit 404 does not generate a speech-based phonetic description that does not use the text of a word, but simply scores the text-based phonetic descriptions proposed by graph generator 400.

The fact that Gupta does not produce a speech-based phonetic description can be seen clearly by removing all of the phonetic descriptions that use text. If this is done, allophone graph generator 400 produces an empty graph because the graph is only populated using letter-to-phoneme rules. (see Col. 5, lines 24-39) This empty graph is provided to graph scorer 404, which is then unable to function since it does not have any phonetic sequences to apply the speech signal against. If Gupta produced a speech-based phonetic description, this would not be true since the speech-based phonetic description would still be present even if the text-based phonetic descriptions were removed.

As noted above, Häb-Umbach also fails to show the selection between a text-based phonetic description and a speech-based phonetic description based in part on the correspondence between each phonetic description and a representation of a speech signal.

Since neither Gupta nor Häb-Umbach select between a text-based phonetic description and a speech-based phonetic description based on the correspondence between the phonetic descriptions and a speech signal, their combination does not show or suggest the invention of claim 12 or claims 13-17, which depend therefrom.

Claim 18

Claim 18 was rejected under 35 U.S.C. § 103(a) as being unpatentable over Gupta in view of Häb-Umbach and further in view of Contolini. Claim 18 depends from claim 12 and thus includes the limitation to selecting between a text-based phonetic description and a speech-based phonetic description based in part on the correspondence between each phonetic description and a

representation of a speech signal. None of the cited references show this limitation.

In particular, Cantolini does not show or suggest selecting between a text-based phonetic description and a speech-based phonetic description, because it does not show or suggest producing a speech-based phonetic description from a speech signal without using the text of the word.

Since none of the cited references select between a text-based phonetic description and a speech-based phonetic description based on a correspondence between the phonetic descriptions and a speech signal, the combination of these references does not show or suggest the invention of claim 18.

Claims 19-21

Claims 19-21 were rejected under 35 U.S.C. § 103(a) as being obvious from Schulze (U.S. Patent No. 6,167,369) in view of Gupta.

Schulze describes a system for determining the language of a document. To do this, Schulze generates a set of trigram models for each language, where each trigram model provides the probability of a character trigram in the language. An input text is then divided into trigrams. The trigrams for the input text are scored using the models for each language to generate a total score for each language. Schulze does not show or suggest syllable-like units or forming n-grams of syllable-like units.

Independent claim 19 provides a speech recognition system with a language model that is trained through a series of steps that include breaking each word in a dictionary into syllable-like units and for each word, grouping the syllable-like units into n-grams. The total number of n-gram occurrences in the dictionary is counted and for each n-gram, the total number of occurrences of the particular n-gram is divided by the total number of n-gram occurrences in the dictionary to form a language model probability for the n-gram.

The combination of Schulze and Gupta does not show or suggest the invention of claim 19. In particular, neither reference shows or suggests grouping syllable-like units found in dictionary words into n-grams.

In the Office Action, it was asserted that Schulze shows grouping syllable-like units from dictionary words into n-grams at column 1, line 29. Applicants respectfully dispute this assertion.

The cited section of Schulze discusses dividing an input sentence into individual character trigrams. It does not mention syllable-like units or forming n-grams from syllable-like units. Furthermore, it would not be obvious to use syllable-like units with Schulze. One goal of the Schulze system is to be able to identify the language of short text segments. If larger units were used instead of individual characters, there would be fewer n-gram probabilities calculated for short text segments thereby making it more difficult to identify the language of the text.

In the Office Action, it was asserted that the sub-word units of Gupta correspond to a syllable-like unit. Thus, the rejection appears to be based on substituting the sub-word units of Gupta in the technique described by Schulze.

However, those skilled in the art would not make such a substitution. Under Schulze, the language of the text is unknown. Because of this, it would be very difficult and in some cases may be impossible to divide the words into syllable-like units. In fact, for some languages in Schulze, the text can not even be divided into words. (See Schulze Col. 15, lines 50-53). Thus, those skilled in the art would not apply the sub-words of Gupta to Schulze as suggested by the Examiner. As such, claim 19 and claims 20 and 21, which depend therefrom, are patentable over the combination of Gupta and Schulze.

Claims 20 and 21 are additionally patentable over Schulze and Gupta. In claim 20, the dictionary words are broken

into syllable-like units by preferring syllable-like units that occur more frequently in the dictionary than other syllable-like units. Neither Schulze nor Gupta show or suggest this additional limitation.

In the Office Action, it was asserted that Schulze showed preferring syllable-like units that occur more often at column 12, lines 35-37. However, the cited section does not discuss syllable-like units or breaking a word into speech units by giving preference to certain speech units. Instead, the cited section states that trigrams with low frequency counts are discarded from a trigram array.

The difference is that Schulze discards trigrams after it has broken the word into trigrams, whereas claims 20 and 21 use preferences for certain speech units during the breaking of words into syllable-like units.

Since the rules in Schulze for identifying trigrams do not show a preference being applied so that one trigram is preferred over another during trigram identification, the cited section of Schulze cannot show or suggest preferring syllable-like units that occur more often in a dictionary over other syllable-like units when dividing words into syllable-like units.

As such, the combination of Gupta and Schulze does not show or suggest the invention of claims 20 and 21.

Conclusion

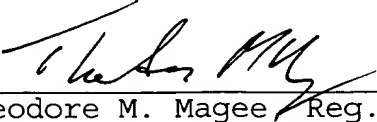
In light of the above remarks, claims 1-21 are patentable over the cited art. Reconsideration and allowance of the claims is respectfully requested.

The Director is authorized to charge any fee deficiency required by this paper or credit any overpayment to Deposit Account No. 23-1123.

Respectfully submitted,

WESTMAN, CHAMPLIN & KELLY, P.A.

By:


Theodore M. Magee, Reg. No. 39,758
Suite 1600 - International Centre
900 Second Avenue South
Minneapolis, Minnesota 55402-3319
Phone: (612) 334-3222 Fax: (612) 334-3312

TMM:sew